Improving Discrimination of Ventricular Tachycardia and Ventricular Fibrillation Using Classifier Ensembles and High Dimensional Representations

Yaqub Alwan, Zoran Cvetković, *Senior Member, IEEE,* Michael J. Curtis

***Abstract*—Differentiating between ventricular tachycardia and ventricular fibrillation is a difficult problem, and highly relevant for clinical practice and translational research. While having low false arrhythmia alarm rates, previous approaches were found to be inadequate for discriminating between these two categories. For improving their discrimination, we introduce the use of high-dimensional feature vectors, in particular, magnitude spectra, and classifier ensembles that take into account local information from the electrocardiogram signals. In order to deal with the increased false arrhythmia alarm rate that results from this approach, a hierarchical classification is proposed, which significantly improves the classification sensitivities of ventricular tachycardia and ventricular fibrillation, while at the same time achieving a modest decrease in the false alarm rate.**

***Index Terms*—Cardiac arrhythmias, ventricular tachycardia, ventricular fibrillation, classifier ensembles.**

I. Introduction

ACCORDING to the World Health Organisation data, cardiovascular disease is the leading cause of death in middle and high income countries, and among the top ten causes of death in low income countries [1]. Development of effective drug treatments that may prevent cardiac arrhythmia is therefore a high-priority challenge for modern pharmacology. For the development of such treatments it is crucial to have a clear understanding of what distinguishes different forms of arrhythmia, and based on that, establish their precise definitions. However, it is evident that although unequivocal ventricular fibrillation (VF), sustained and lethal, is incontestable in electrocardiogram (ECG) recordings, clinicians differ fundamentally about the diagnosis and appellation of transient polymorphic ventricular tachyarrhythmias, with experts in a landmark report unable to agree on whether VF, polymorphic ventricular tachycardia (VT) or torsade de pointes best described a range of human tachyarrhythmias in a blinded test of ECG records [2]. Given that mechanisms of these tachyarrhythmias may differ [3] and responses to drugs may vary from benefit to proarrhythmia, depending on the type [4], errors in diagnosis due to unequivocal appellations are potentially hazardous. From a therapeutic point of view being able to differentiate between VF and VT is very important since they respond to interventions differently and VF is often lethal, while VT is often not. To allow preclinical research to be translatable, guidance was proposed, and recently updated, on differentiation between VF, including brief and transient VF, and other polymorphic ventricular tachyarrhythmias [5], [6]. The guidance, however, is not readily transformed into an algorithm for automatic rhythm classification. Therefore, in the present study we build upon existing procedures for ventricular tachyarrhythmia classification with a focus on improving discrimination between VT and VF.

There have been many studies into the topic of differentiating sinus rhythm (SR), including all rhythms that are not VT or VF, from VF or ventricular arrhythmias, however few studies attempt to differentiate VT from VF, and even fewer perform three class classification between SR, VT and VF [7]–[10].

Schemes for detecting ventricular arrhythmias, but with no focus on separating VT from VF include measuring the leakage from an adaptive bandstop filter [11], counting the number of boxes filled in variations of a phase space representation [12], [13], measuring the sample entropy of the ECG [14] and comparing with landmark features of some predefined templates [15].

Approaches which do try to make decisions between VT and VF, include sequential hypothesis testing on a count of threshold crossings [9], measuring the number of islands in a time-scale representation [16], the area occupied in the bispectral representation [8], standard deviation between peak amplitudes and peak distances [17], the spectral entropy and energy of the first empirical mode decomposition component [10], the time between cardiac deflections [7], and computing various statistics on a bandpass filtered version of the signal [18].

Many of these studies suffer from one or more of the following experimental drawbacks:

- Introducing a separate, third category for VT and VF examples which are hard to differentiate [7], [16];
- Using the same samples for training and testing;
- Developing an ad-hoc decision scheme rather than using well understood and statistically motivated decision-making algorithms for classification;
- Using hand selected and annotated data, often not available for scrutiny.

All previous studies also have in common the use of very low dimensional, mainly heuristic, feature vectors to represent the data existing originally in high dimensional space, which could potentially remove the information needed to discriminate between different arrhythmias.

Recent papers reporting results of comprehensive comparative studies [15], [19], [20], have covered many of these schemes, or at least their feature spaces, and also many others not mentioned here. None of these studies specifically focus on differentiating VT from VF, however it has been noted that separating the two is very challenging [20], [21].

There is therefore a need for systematic investigation of cardiac arrhythmia classification using representations which involve minimal information reduction and well established and understood classification methods. This is provided in part by recent studies where classification between SR and ventricular arrhytmias [19], [20], or between non-VF and VF [20] is considered using support vector machines (SVMs). However, these studies do not consider classification between all three groups SR, VT and VF, and the dimension of the representation space is still small, at a maximum of 14. In this study, we focus on classification between SR, VT, and VF, using as few heuristics as possible, in higher dimensional feature spaces provided by Fourier magnitude spectra of ECG waveforms. We further introduce local context ensembles, composed of consecutive ECG segments, to gain additional classification accuracy. Classification experiments using SVMs show that while high-dimensional magnitude spectra achieve significant improvements in discriminating between VT and VF, state-of-the-art low-dimensional features exhibit lower false arrhythmia alarm rate. Finally, a hierarchical approach to ECG classification is proposed, that uses the low-dimensional features for making the SR vs non-SR decisions, and then magnitude spectra for the discrimination between VT and VF. Such a hierarchical classifier achieves improvements in sensitivities for all three classes.

The paper is organised as follows. Section II sets out the basic framework of the problem and discusses in some detail the prior art used for comparison. Section III provides details of the classification framework used. Experimental procedure and results are reported in Section IV. Section V summarises the paper and draws conclusions.

II. General Considerations

Given a segment $\mathbf{x}$ of discretised ECG signal,

$$\mathbf{x} = \{x[n], n_1 \leq n \leq n_2\} \ ,$$

we wish to be able to classify it as SR, VT or VF. The classification algorithm is in general a function C of the following form:

$$C(\mathbf{x}) = f(T(\mathbf{x})) \ , \tag{1}$$

where T is some transform function whose output is a vector of features, while f is some decision function. In previous studies, most often some heuristic transforms T, combined with empirical decision functions f have been used. Recently SVMs were introduced, as a formal decision making framework, in two studies which provide a comparative analysis of previously proposed transforms and explore improving classification accuracy by combining highly ranked features [19], [20]. A subset of highest ranked or common transformations considered in [19], [20] are used as a reference in this work.

Here we also use SVMs for assigning class labels due to their good generalisation ability. We will combine SVMs with error-correcting output code methods [22] in order to perform three-way classification, and ensemble methods [23], [24] in order to improve results by aggregating decisions made on consecutive ECG segments. In the domain of transformations T, we introduce magnitude spectra of ECG segments. The rationale behind this set of features is that they preserve most of the information in ECG signals, and thus might aid discrimination between VT and VF, but removal of the phase makes the transformation shift-invariant. Preliminary investigations demonstrated a considerable advantage of using Fourier magnitude spectra instead of underlying ECG waveforms, *i.e.* the complete information [21], [25].

A. Reference Transformations

The transformations proposed in the previous work, which will be used here as a reference are the following:

1) *VF filter leakage* [11], which is the residual energy obtained after applying an adaptive bandstop filter centered at the mean frequency of the considered ECG segment $\mathbf{x}$.

2) *Count 2* [18], which is obtained by applying a narrow bandpass filter to $\mathbf{x}$ and then counting the number of samples of the output $\hat{\mathbf{x}}$ satisfying $\text{mean}(|\hat{\mathbf{x}}|) \leq |\hat{\mathbf{x}}| \leq \max(|\hat{\mathbf{x}}|)$.
3) *Threshold crossing sample count* [26], which is an improvement to the threshold crossing interval transformation [9], obtained by counting number of samples above the absolute value of an adaptive threshold.
4) The *sample entropy* of $\mathbf{x}$ [14], computed in a standard way.
5) *Spectral parameters* m and $A2$ [27]. To compute these two parameters, first the discrete Fourier transform of $\mathbf{x}$ is found and F, the frequency with the largest amplitude between 0.5 Hz and 9 Hz, is identified. Then m and $A2$ are obtained as

$$m = \frac{\sum_i A_i f_i}{F \sum_i A_i} \ ,$$

$$A2 = \sum_{i:0.7F \leq f_i \leq 1.4F} A_i \ ,$$

where f_i is the i^{th} frequency in the spectrum, and A_i is the absolute value of the discrete Fourier transform at f_i.
6) *PST* [12] and *PSH* [13] phase space parameters. To compute these two parameters, phase spaces of $\mathbf{x}$ are formed, one using a time delay [12] and one using the Hilbert transform [13]. Each phase space is then quantised and the number of unique value-pairs is counted, giving *PST* and *PSH* parameters.

Each of the above features were originally designed with different observation lengths in mind, however they are all easily extended to arbitrary observation lengths. As one reference set of features we use VF filter leakage and Count 2 parameters combined, as recommended in [19]; we will refer to this representation as Heur2. The other reference set of features which we will use is the full set of the parameters described in the above; we will refer to this representation as Heur8. Note that the Heur8 feature set is composed of the two parameters which achieved the highest performance rank in [19] and the six parameters which achieved the highest performance rank in VF vs non-VF classification in [20].

The range of observation lengths considered includes both 2 s and 8 s, which are observation lengths used in the two reference studies, [19] and [20], respectively. Note that all of these previously proposed feature sets are low-dimensional, and that even their combinations, as considered in [19], [20] do not exceed 14 features. The dimension of Fourier magnitude spectra, as proposed here, even in the case of 1 s ECG segments is 50; this additional information, as it will be shown later, will provide an advantage in discriminating between VT and VF.

B. Observation length

One of the important issues that needs to be investigated is the appropriate observation length for reliable classification of cardiac arrhythmias. However, apart from the study in [19] which did not consider VT and VF discrimination, there is no comprehensive study into the impact of window length.

Most previous investigations select a single observation length for analysis, and they are often considerably long, ranging from 5 s to 14 s. Physiological considerations, on the other hand, suggest that shorter ECG segments should suffice. VT is considered to occur if 4 or more consecutive QRS complexes precede their corresponding P-wave, independent of the rate [5]. If we also assume a base heart rate of 60 beats per minute in humans, then a 2 s window should be sufficient to capture 2 normal beats. Thus, at least 3 premature QRS complexes would occur in the same 2 s interval, and since VT is usually accompanied by an increased heart rate, quite often 2 s should also be sufficient to capture 4 or more premature QRS complexes. Since it captures sufficient number of QRS complexes, or lack thereof in the case of VF, a 2 s analysis window should be sufficient for good discrimination, and possibly even a 1 s observation window. In fact, our preliminary study revealed that classification performance does not improve considerably with the increase of observation window from 1 s to 4 s [25]. Thus, in Section III-C we consider ensemble methods for improving classification accuracy by combining decisions made on consecutive short ECG segments taken over longer observation periods.

III. Classification using support vector machines

Given a set of training data $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with corresponding class labels $(y_1, \ldots, y_p)$, $y_i \in \{+1, -1\}$, an SVM aims to find a decision surface which jointly maximizes the margin between the two classes and minimizes the misclassification error on the training set. When the classes are linearly separable, these surfaces are linear and have the form

$$f(\mathbf{x}) = \sum_i a_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b = 0, \ a_i \in \mathbb{R}^+ \qquad (2)$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^n$, while the Lagrange multipliers a_i and the bias b are optimized by the training algorithm. Non-linear separators between two

classes are created by means of non-linear kernel functions $K(\cdot,\cdot)$. These functions compute inner products in higher dimensional spaces, without explicitly performing the mapping, where the data could potentially be linearly separable. Analogously to the linearly separable case, the decision surface is constructed according to

$$f(\mathbf{x}) = \sum_i a_i y_i K(\mathbf{x}, \mathbf{x}_i) + b = 0 \tag{3}$$

and the class label of a test vector $\mathbf{x}$ is predicted to be the sign of the score function evaluated at $\mathbf{x}$:

$$C(\mathbf{x}) = \mathrm{sgn}(f(\mathbf{x})) \ .$$

A commonly used kernel function is the radial basis function (RBF) kernel given by

$$K_r(x,y) = e^{-\gamma||x-y||^2}, \gamma \in \mathbb{R}^+ \tag{4}$$

where γ is optimised using a grid search. This is the only kernel considered in other studies [19], [20]. Our preliminary investigations which included also polynomial kernels found that better results are indeed obtained with the RBF kernel [25], so all results reported in this paper are obtained with this kernel.

A. Optimising SVM parameters

SVMs involve some free parameters that need to be optimised, to achieve good classification performance on unseen examples. For SVMs using the RBF kernel, these parameters are:

$\mathcal{C}$: Trade-off between margin width and misclassified examples from the training data;

γ: RBF kernel parameter which controls the width of the Gaussian function. Small values of γ lead to increasing flexibility of the decision boundary, while large values of γ make the boundary less flexible.

These parameters were tuned by means of a grid search. $\mathcal{C}$ was searched over the range $\{10^{-5}, \ldots, 10^{0}\}$, and γ was searched over the range

$$\gamma_{search} = 10^N, N \in [\gamma_{start} - 2, \gamma_{start} + 2] \ , \tag{5}$$

where γ_{start} is given as

$$\gamma_{start} = -\log_{10} D_{mean} \ , \tag{6}$$

while D_{mean} is the mean norm of training vectors.

When optimising these parameters, five fold cross validation was used to find the best pair $(\gamma, \mathcal{C})$ over the entire training set.

B. Multiclass classification using SVMs

For multiclass discrimination, binary SVM classifiers are combined via predefined error-correcting output code methods [22], [28]. To achieve this, N binary classifiers are trained to distinguish between M classes using a coding matrix $\mathbf{W}_{M\times N}$, with elements $w_{mn} \in \{0, 1, -1\}$. Classifier n is trained only on data of classes m for which $w_{mn} \neq 0$, with $\mathrm{sgn}(w_{mn})$ as the class label. Then, the class assignment rule is given by

$$C(\mathbf{x}) = \arg\min_m \sum_{n=1}^{N} \chi(w_{mn} f_n(\mathbf{x})) \ , \tag{7}$$

where $f_n(\mathbf{x})$ is the output of the n^{th} classifier and χ is some loss function.

The error-correcting capability of a code is commensurate with the minimum Hamming distance between the rows of a coding matrix; if this minimum distance is δ, then the decoding process will be able to correct any $\lfloor \frac{\delta-1}{2} \rfloor$ errors [28]. For the three class problem, we consider all *one-vs-one* (pairwise) and all *one-vs-all* binary classifiers, which makes a total of six binary classifiers. In the case of three classes, this exhausts all possible binary classifiers. The corresponding coding matrix in this case thus has the form

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & -1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 1 \\ -1 & 1 & 1 & -1 & 0 & -1 \end{bmatrix} \ . \tag{8}$$

A number of choices for loss functions exist, including hinge: $\chi(z) = \max(1 - z, 0)$, Hamming: $\chi(z) = [1 - \mathrm{sgn}(z)]/2$, exponential: $\chi(z) = e^{-z}$, and linear: $\chi(z) = -z$ function.

C. Local context ensembles

Ensembles of classifiers can often be combined to improve classification accuracy [23], [24]. Since there are diminishing returns to performing classification directly on increasing observation lengths of ECG [19], [21], [25], we propose to combine outputs of binary SVM classifiers applied to ECG segments of a given length, taken with incremental shifts with respect to each other, e.g. 2 s segments taken over 4 s intervals with 0.5 s shifts, to form ensembles of 5 decision values. One way in which these outputs can be combined is via majority voting, while more generally one can combine decision values of individual classifiers in the ensemble. This can be realised by forming vectors $\mathbf{f}_n$ of decision values corresponding to K consecutive ECG segments $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$,

$$\mathbf{f}_n = (f_n(\mathbf{x}_1) \ f_n(\mathbf{x}_2) \ \ldots \ f_n(\mathbf{x}_K)), \tag{9}$$

and using an aggregation function A. Combined with (7), this gives the decision rule as

$$C(\mathbf{x}) = \arg\min_{m} \sum_{n=1}^{N} \chi(w_{mn} A(\mathbf{f}_n)) \ . \tag{10}$$

Possible choices for $A(\mathbf{f})$ include but are not limited to (i) $A(\mathbf{f}) = \text{mean}(\mathbf{f})$ (ii) $A(\mathbf{f}) = \text{median}(\mathbf{f})$ (iii) majority voting: $A(\mathbf{f}) = \text{mode}(\text{sgn}(\mathbf{f}))$, and (iv) maximum absolute value: $A(\mathbf{f}) = \mathbf{f}_{\ \arg\max_k |\mathbf{f}|}$.

We refer to this type of decoding as local context ensembles (LCEs). An important thing to note about LCE decoding is that it can be performed forwards, or backwards. In case that it is performed forwards, a latency is incurred equal to the total duration used for decoding. Thus, in the context of a real-time analyser, a decision for a specific point in time when using decoding over 8 s is not made until 8 s of ECG is acquired, even if the base observation length is only 1 s. On the other hand, when using backwards decoding, the amount of data that needs to be acquired is only the base observation length, which has the potential to reduce the amount of time required to make a decision significantly. As we found no significant difference in performance between forwards and backwards decoding, in this study we report only results obtained with backwards LCE decoding.

D. Hierarchical classification

In case that some classifiers are better at certain tasks than others, and the task can be broken down into a sensible hierarchy, it can be worthwhile considering a hierarchical structure, with different transformations being used for different parts of the decision hierarchy until a final decision is made. We observed that one set of features achieved higher sensitivity for the SR class, while another set of features achieved higher sensitivities for the two types of arrhythmias. Hence, we consider the following hierarchical classification structure:

$$C(\mathbf{x}) = \begin{cases} C_1(\mathbf{x}), & C_1(\mathbf{x}) = \text{SR} \\ C_2(\mathbf{x}), & C_1(\mathbf{x}) \neq \text{SR} \end{cases} \tag{11}$$

Here, C_1 and C_2 can be composed in many ways, by varying f and T. For example, C_2 can be restricted to output only VT or VF, or it may be allowed to make a SR decision. C_1 and C_2 may use different T, or different f, or both.

IV. Experimental Setup and Results

In this section we discuss the details of the experimental procedures, the data used and processing applied, followed by results presentation and discussion.

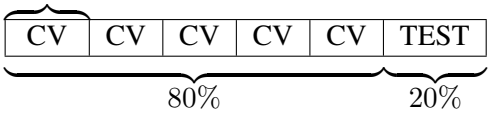


Fig. 1. The ECG records are partitioned into two groups, training, and testing.The training partitions are further split into folds, for cross-validation. After free parameters are estimated with CV sets, these sets are used for a final training pass with optimised parameters, and accuracy is assessed with the final partition labelled TEST.

A. Metrics for assessing model accuracy

Given a set of category labels S, we can define multi-category sensitivity of a particular category s as,

$$sn_s = \frac{TP_s}{TP_s + FN_s}, \ s \in S \ , \tag{12}$$

where TP_s is the amount of category s correctly identified as s, and FN_s is the amount of category s incorrectly assigned to other categories. Then, balanced accuracy, or average sensitivity, can be defined as

$$Acc_{bal} = \frac{1}{|S|} \sum_{s \in S} sn_s \ . \tag{13}$$

B. Methods for estimating predictive power

To estimate the predictive power of a learned model, it is important to emulate the real scenario where unseen data is used for prediction. This usually involves having separate data for model building and assessing accuracy. Two commonly used techniques used are cross validation and bootstrap resampling. Bootstrap resampling divides the data set randomly into two parts, with one used for training and one for testing. This is repeated many times with different randomised partitions. Cross validation partitions the data into N mutually exclusive equally sized sets, where one is left for testing and the remainder left for training. This is repeated with all the sets used as the testing set once. The number of sets is user selectable.

For this study we selected bootstrap resampling with 50 resamples for estimating distributions of generalisation accuracy, and five fold cross validation for tuning free parameters across the training set. The resampling and cross validation process is done per record, rather than per data point, and is detailed in Fig. 1.

C. Data and preprocessing

1) Data sets: Data were taken from Physiobank [29], which maintains a large online repository of various physiological signals, including ECG signals. The databases used from Physiobank were the European ST-T Database [30], the Creighton University Ventricular

Tachyarrhythmia Database [31], the MIT-BIH Arrhythmia Database [32] and the MIT-BIH Malignant Ventricular Arrhythmia Database [33]. We also used the extended American Heart Association Database. Only records from these databases containing examples of VT or VF were used. Any records containing annotations for so-called ventricular flutter were excluded, due to ambiguity about which category these rhythms belong to. This results in 91 out of 98 possible records being used.

2) Preprocessing: In all these databases, 250 Hz sampling rate is used, apart from MIT-BIH Arrhythmia Database (MITDB) where signals are sampled at 360 Hz. It is considered that most of the relevant information is contained in the 40 Hz baseband [34] and that preprocessing with a 30 Hz low pass filter does not affect experimental results [10], [12], [15], [16]. However, based on visual inspection of low-pass filtered data it was decided that 30 Hz cut-off frequency was too low, so 50 Hz low-pass filtering was used, followed by downsampling to 100 Hz. In addition to this, a 0.5 Hz high pass filter was applied to remove wandering baseline [34]. All ECG records were normalised so that the sum of squares of each record is equal to the number of samples in the record, thus making the variance of individual time samples equal to 1.

3) Data balancing: For training and testing of SVMs, we used the GTSVM software [35]. Since this software does not support different costs for the different categories, it was necessary to use balanced training data by randomly undersampling categories with more points (SR, VF). Some preliminary experiments using another SVM package showed that there was no significant difference between using balanced training data or imbalanced training data with different costs per category.

D. Experiments and Results

We conducted three-way classification tasks between SR, VT and VF using three different representation spaces. The Fourier magnitude spectra (Spectra) of ECG windows was used as a high dimensional representation space with minimal information reduction. For comparison purposes, we used the Count2 and VFLEAK features (Heur2) suggested to perform well [19] and all 8 features described in II-A (Heur8).

These classification tasks were performed using non-overlapping segments for training and testing, and also overlapping segments with 0.25 s shifts for training in the case of LCEs.

Fig. 2 shows all the different variations on LCE decoding methods, and how on average, they improve upon the base classification accuracy obtained without LCE decoding. Backwards LCE decoding is performed using data trained on base observation lengths of 1 s, 2 s and 4 s and a number of previous segments, up to 8 s in the past. The parameters were varied through number of previous segments (increasing, up to a maximum of 8 s, with 1 s increments), shift amount (0.25, 0.5 s), χ (hinge loss, linear loss, exponential loss, and Hamming loss), and A (mean, median, majority vote, maximum

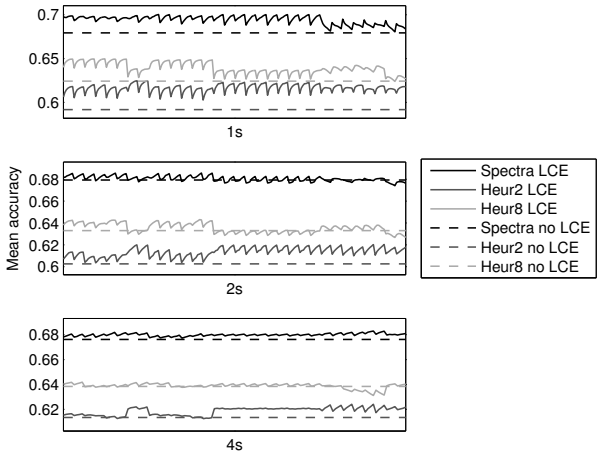


Fig. 2. LCE decoding vs baseline accuracy. The accuracy shown is the mean accuracy across all bootstrap resamples. The dashed line shows baseline classification accuracy obtained without LCE decoding. The LCE decoding method for each representation space and base observation length is varied in the order: number of consecutive segments, from base length plus 1 s to 8 s, offset of consecutive segments (0.25 s or 0.5 s), variations of χ (hinge loss, linear loss, exponential loss, and Hamming loss) and A (mean, median, majority vote, maximum absolute value).

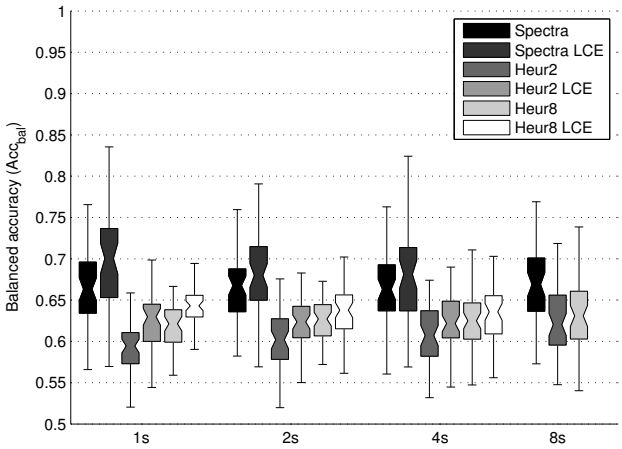


Fig. 3. Boxplots showing distributions of average sensitivity across all categories and bootstrap resamples as computed using (13) for each resample. Distributions are shown for non overlapping observation windows and the best LCE as taken from Fig. 2. Results are shown for all three representation spaces, using base window lengths of 1 s, 2 s, 4 s and 8 s.

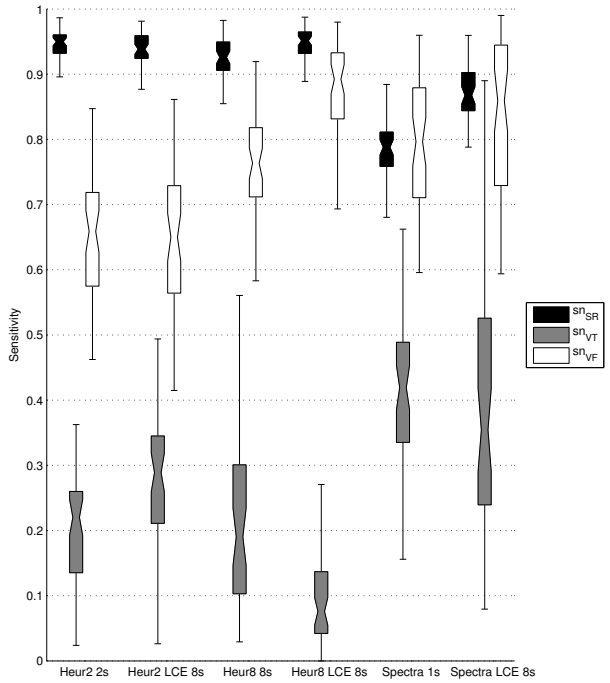
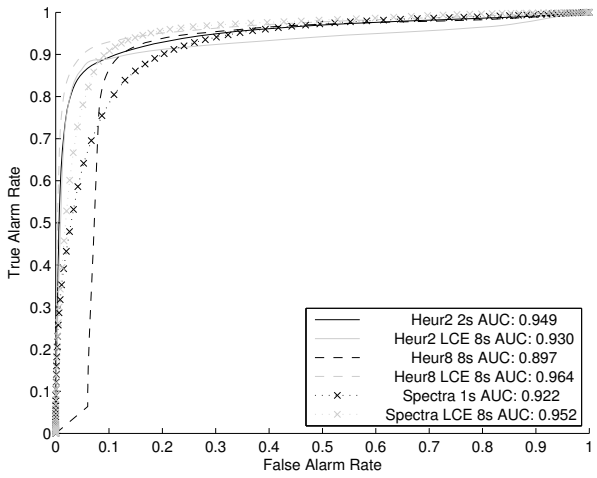


Fig. 4. Boxplots showing distributions of sensitivities for each category across all bootstrap resamples. Groups of sensitivites are shown for all three representation spaces, using base window lengths as specified in the original studies, or in the case of Spectra, the best performing window size. These are also shown alongside the best LCE in terms of average sensitivity for the given representation space.

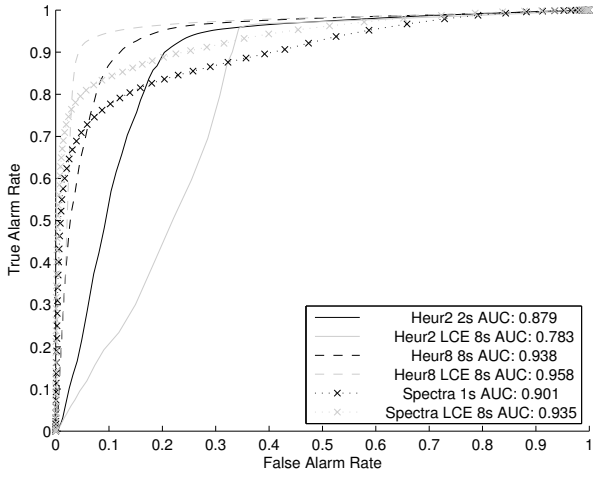
TABLE I

AVERAGE CONFUSION MATRICES FOR EACH REPRESENTATION SHOWN IN FIG. 4. ROWS ARE THE GROUND TRUTHS, AND COLUMNS ARE THE DIAGNOSES MADE. THE FINAL COLUMN SHOWS ACCURACY (MEAN ALONG THE DIAGONAL).

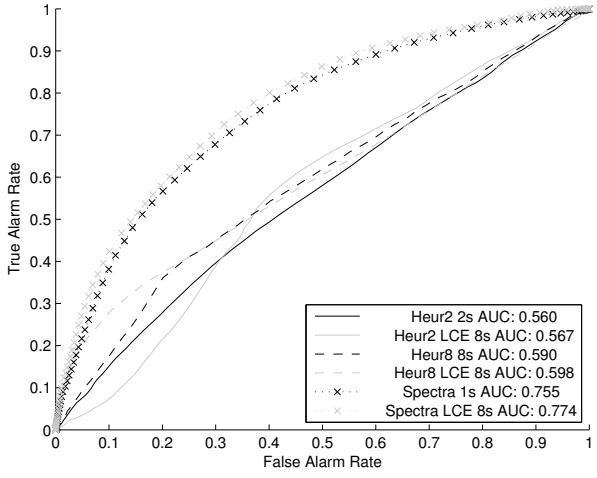
Method	**SR%**	**VT%**	**VF%**	**ACC%**
Heur2 2s	94.5	2.5	3.0	
	25.0	20.0	55.0	60.1
	10.0	24.4	65.6	
Heur2 LCE 8s	93.9	4.1	1.9	
	21.6	27.4	51.0	62.0
	8.0	27.2	64.8	
Heur8 8s	92.3	5.1	2.6	
	15.7	23.8	60.5	64.1
	7.5	16.3	76.1	
Heur8 LCE 8s	94.4	3.0	2.6	
	21.3	12.6	66.2	65.0
	4.0	8.0	88.0	
Spectra 1s	78.2	7.0	14.7	
	19.5	43.7	36.8	67.0
	4.9	15.9	79.2	
Spectra LCE 8s	87.1	4.0	8.8	
	21.9	39.4	38.7	70.0
	2.6	13.8	83.6	



(a)



(b)



(c)

Fig. 5. Average receiver operating characteristic curves computed across all bootstrap resamples from the testing records. AUC is area under the curve. (a) SR vs arrhythmia classifier, (b) non-VF vs VF and (c) VT vs VF. These are shown for the same representation spaces as in Fig. 4 and TABLE I

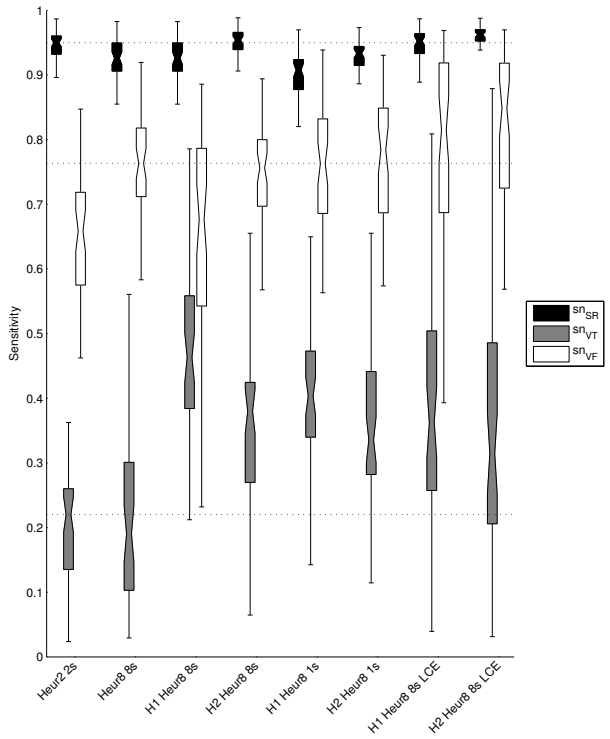


Fig. 6. Boxplots showing sensitivity distributions for each category across all bootstrap resamples for hierarchical decision making compared with results from Heur2 and Heur8 representations. The representation space named for each classifier is the first decision representation space, and the second decision is always made by a Spectra classifier which matches the observation length and decoding parameters of the first decision classifier. H1 means the hierarchy did not allow the Spectra classifier to make a SR decision, H2 means that the spectra classifier could decide SR. The dashed lines correspond to the best median for each category from either Heur2 or Heur8.

absolute value), varied in this order. As can be seen from the sawtooth shape of the graphs, adding more consecutive segments improved the accuracy the most, although choice of A appeared to have a significant impact in some cases. The most significant improvement was obtained when LCEs were composed using 1 s base units, with less gains obtained from LCE decoding using longer base observation intervals. From here on, reported metrics for LCE decoding methods is shown only for the best variation across all parameters for each representation space and base observation length.

Fig. 3 uses boxplots to show the distributions of average sensitivities across all bootstrap resamples. These distributions are shown for all representations and observation length combinations, both for non-overlapping testing, and for the best LCE decoding method among all the decoding variations trained using overlapping observations. Classification using the Spectra representation had a significantly higher median than methods using other representation spaces, even without LCE decoding. Significant improvement to the median classification accuracy was observed when using LCE decoding for all representation spaces. However, neither Heur2 or Heur8 representation spaces with LCE decoding performed better than the Spectra representation without LCE decoding. The highest average accuracy is obtained when using LCE decoding on 1 s base observations.

In order to understand how Spectra improved the average sensitivity, Fig. 4 shows distributions of sensitivities per category across all bootstrap resamples. These are shown for Heur2 and Heur8 using the same observation lengths as in the original studies, 2 s and 8 s, respectively, and also for Spectra trained using 1 s observation lengths. Each representation is also shown with its best LCE decoding method, formed over 8 s using 2 s observations for Heur2, and over 8 s using 1 s observations for Heur8 and Spectra. TABLE I shows the average of normalised confusion matrices over all bootstrap resamples for the same representation spaces. Spectra classified with 1 s observation windows obtained considerably higher VT and VF sensitivities than the Heur2 and Heur8 representations as reported with their original observation lengths. However, the SR sensitivity was reduced when compared to these two representations.

Since multi-class classifiers are not amenable to receiver operating characteristic analysis, and since multi-class SVMs are formed using binary classifiers, in Fig. 5 we show receiver operating characteristic curves for the individual binary classifiers SR vs arrhythmia (Fig. 5a), non-VF vs VF (Fig. 5b) and VT vs VF (Fig. 5c). These are shown for the same representation spaces as in Fig. 4. We can see that the best receiver operating characteristic, according to the area under the curve, for non-VF vs VF and SR vs arrhythmias was the Heur8 LCE classifier. However, for the VT vs VF case, the Heur2 and Heur8 classifiers were operating at the no discrimination point. In this case, only the Spectra classifier was capable of making decisions considerably better than the level of guessing.

The previous results motivate a hierarchical classifier architecture. Since the Spectra representation space gave better results for VT vs VF, and Heur2 and Heur8 representations had lower false arrhythmia alarm rates, for hierarchical classification the first decision is made by a Heur2 or Heur8 representation classifier. Then, if an arrhythmia decision is made by the first classifier, the final decision is delegated to a Spectra classifier with matching window length and decoding parameters. We tested two variations of this scheme. The first variation, referred to as H1, allows the Spectra classifier to make only VT or VF decisions. The second variation, referred to as H2, allows the Spectra classifier to make a full

range of decisions, SR, VT or VF. The second variant is motivated by the fact that classifications with the Spectra representation assigned fewer examples of arrhythmias to SR allowing for corrections to arrhythmia diagnoses made by the first classifier.

In Fig. 6 we show the distributions of sensitivities of each category. These are shown for a variety of hierarchical constructions. This information is also summarised in TABLE II. Hierarchical classifiers obtained using the Heur2 representation performed worse than those constructed using Heur8 representation, so we only show results obtained with the latter representation. The inferior performance of hierarchical classifiers which use the Heur2 features can be explained by the confusion matrices in TABLE I. It can be seen from these confusion matrices that many more VF are assigned to the SR category on average for the Heur2 representations. This causes incorrect SR decisions that cannot be corrected by the Spectra classifier, due to the nature of the hierarchical construction. It can be seen that all of the constructions shown improved upon the VT sensitivity compared with Heur2 and Heur8 representations, and either equalled or improved VF sensitivities. Highest VT sensitivities were obtained when using a H1 hierarchy, however using a H2 hierarchy slightly decreased VT sensitivity, but allowed the SR sensitivity to exceed that of the Heur2 and Heur8 representations as presented by the original studies. Building either a H1 or H2 hierarchy using the Heur8 representation and LCE decoding over 8 s using 1 s windows improves the sensitivities of all categories compared to classifying with just Heur8 or Heur2 representations.

As it can be seen from TABLE II the improvement in median balanced accuracy achieved by hierarchical classification using Spectra and LCEs is from 60.2% (Heur2) and 63.1% (Heur8) to $70.3\% - 70.5\%$. The improvement to VT and VF sensitivities is considerable; in the case of VT the improvement is from 22.0% (Heur2) and 19.0% (Heur8) to $31.5 - 36.2\%$, while for VF the improvement is from 65.9% (Heur2) and 76.3% (Heur8) to $81.5 - 84.8\%$.

TABLE II
MEDIAN SENSITIVITIES OF EACH CATEGORY, AND MEDIAN BALANCED ACCURACY, SHOWN FOR SPECTRA, HEUR2 AND HEUR8 REPRESENTATIONS WITHOUT ANY LCE OR HIERARCHY. ALSO SHOWN FOR SOME BEST PERFORMING HIERARCHICAL CLASSIFIERS, WITH AND WITHOUT LCE DECODING

Method	**SR%**	**VT%**	**VF%**	Acc_{bal}**%**
Heur2 2s	95.0	22.0	65.9	60.2
Heur8 8s	92.6	19.0	76.3	63.1
H1 Heur8 8s	92.6	46.3	67.7	68.7
H2 Heur8 8s	95.4	38.0	75.6	69.0
H1 Heur8 1s	90.8	40.3	76.3	69.7
H2 Heur8 1s	93.3	33.6	78.4	69.0
H1 Heur8 8s LCE	95.2	36.2	81.5	70.5
H2 Heur8 8s LCE	96.2	31.5	84.8	70.3

V. CONCLUSIONS

We conducted an investigation into the classification performance and tradeoffs when considering the clinically important SR vs VT vs VF scenario. Representation spaces from previous studies were considered and we found that these representation spaces have poor ability to discriminate between VT and VF. In order to increase VT and VF sensitivities, we introduced Fourier magnitude spectra as a representation space. This successfully increased VF and VT sensitivities, but also increased the false arrhythmia alarm rate. We introduced local context ensemble methods which take into account recent previous information from the ECG, up to a total observation length that is similar to what is used in previous studies, and found it improved the median classification accuracy significantly, although again there was not an improvement for all sensitivities. In addition, using backwards LCE decoding allows for diagnoses to be made with smaller latency than simply classifying using a full observation of equivalent length. We then proposed hierarchical classification constructed from classifiers with different representation spaces in order to improve arrhythmia sensitivities without increasing the false arrhythmia alarm rate. This approach actually reduced false arrhythmia alarm rate while increasing the sensitivity of detection of both VF and VT when compared to previous studies.

REFERENCES

[1] W. H. Organization, "Who — the top 10 causes of death," http://www.who.int/mediacentre/factsheets/fs310/en/index.html, 2012, [Online; accessed 8-November-2012].

[2] R. Clayton, A. Murray, P. Higham, and R. Campbell, "Self-terminating ventricular tachyarrhythmias - a diagnostic dilemma?" *The Lancet*, vol. 341, no. 8837, pp. 93–95, 1993.

[3] H. Clements-Jewery, D. Hearse, and M. Curtis, "Phase 2 ventricular arrhythmias in acute myocardial infarction: a neglected target for therapeutic antiarrhythmic drug development and for safety pharmacology evaluation," *Br. J. Pharmacol.*, vol. 145, no. 5, pp. 551–564, Jul 2005.

[4] M. Chang, E. de Lange, G. Calmettes, A. Garfinkel, Z. Qu, and J. Weiss, "Pro- and antiarrhythmic effects of ATP-sensitive potassium current activation on reentry during early afterdepolarization-mediated arrhythmias," *Heart Rhythm*, vol. 10, no. 4, pp. 575–582, Apr 2013.

[5] M. J. A. Walker, M. J. Curtis, D. J. Hearse, R. W. F. Campbell, M. J. Janse, D. M. Yellon, S. M. Cobbe, S. J. Coker, J. B. Harness, D. W. G. Harron, A. J. Higgins, D. G. Julian, M. J.

Lab, A. S. Manning, B. J. Northover, J. R. Parratt, R. A. Riemersma, E. Riva, D. C. Russell, D. J. Sheridan, E. Winslow, and B. Woodward, "The lambeth conventions: guidelines for the study of arrhythmias in ischaemia, infarction, and reperfusion," *Cardiovascular Research*, vol. 22, no. 7, pp. 447–455, 1988.

[6] M. J. Curtis, J. C. Hancox, A. Farkas, C. L. Wainwright, C. L. Stables, D. A. Saint, H. Clements-Jewery, P. D. Lambiase, G. E. Billman, M. J. Janse, M. K. Pugsley, G. A. Ng, D. M. Roden, A. J. Camm, and M. J. Walker, "The lambeth conventions (ii): Guidelines for the study of animal and human ventricular and supraventricular arrhythmias," *Pharmacology & Therapeutics*, vol. 139, no. 2, pp. 213–248, 2013.

[7] W. Olson, D. Peterson, L. Ruetz, B. Gunderson, and M. Fang-Yen, "Discrimination of fast ventricular tachycardia from ventricular fibrillation and slow ventricular tachycardia for an implantable pacer-cardioverter-defibrillator," in *Computers in Cardiology*, September 1993, pp. 835–838.

[8] L. Khadra, A. Al-Fahoum, and S. Binajjaj, "A quantitative analysis approach for cardiac arrhythmia classification using higher order spectral techniques," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 11, pp. 1840–1845, Nov 2005.

[9] N. Thakor, Y.-S. Zhu, and K.-Y. Pan, "Ventricular tachycardia and fibrillation detection by a sequential hypothesis testing algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 9, pp. 837–843, September 1990.

[10] B. Bai and Y. Wang, "Ventricular fibrillation detection based on empirical mode decomposition," in *5th International Conference on Bioinformatics and Biomedical Engineering*, May 2011, pp. 1–4.

[11] S. Kuo and R. Dillman, "Computer detection of ventricular fibrillation," in *Computers in Cardiology*, 1978, pp. 347–349.

[12] A. Amann, R. Tratnig, and K. Unterkofler, "Detecting ventricular fibrillation by time-delay methods," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 1, pp. 174–177, Jan 2007.

[13] ——, "A new ventricular fibrillation detection algorithm for automated external defibrillators," in *Computers in Cardiology*, Sept 2005, pp. 559–562.

[14] H. Li, W. Han, C. Hu, and M.-H. Meng, "Detecting ventricular fibrillation by fast algorithm of dynamic sample entropy," in *IEEE International Conference on Robotics and Biomimetics*, Dec 2009, pp. 1105–1110.

[15] A. Amann, R. Tratnig, and K. Unterkofler, "Reliability of old and new ventricular fibrillation detection algorithms for automated external defibrillators," *BioMedical Engineering OnLine*, vol. 4, no. 1, p. 60, 2005.

[16] K. Balasundaram, S. Masse, K. Nair, T. Farid, K. Nanthakumar, and K. Umapathy, "Wavelet-based features for characterizing ventricular arrhythmias in optimizing treatment options," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, September 2011, pp. 969–972.

[17] J. Ruiz, E. Aramendi, S. Ruiz de Gauna, A. Lazkano, L. Leturiondo, and J. Gutierrez, "Distinction of ventricular fibrillation and ventricular tachycardia using cross correlation," in *Computers in Cardiology*, September 2003, pp. 729–732.

[18] I. Jekova and V. Krasteva, "Real time detection of ventricular fibrillation and tachycardia," *Physiological measurement*, vol. 25, no. 5, pp. 1167–1178, Oct 2004.

[19] Q. Li, C. Rajagopalan, and G. Clifford, "Ventricular fibrillation and tachycardia classification using a machine learning approach," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1607–1613, June 2014.

[20] F. Alonso-Atienza, E. Morgado, L. Fernandez-Martinez, A. Garcia-Alberola, and J. Rojo-Alvarez, "Detection of life-threatening arrhythmias using feature selection and support vector machines," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 832–840, March 2014.

[21] Y. Alwan, Z. Cvetković, and M. J. Curtis, "High-dimensional discriminant analysis of human cardiac arrhythmias," in *European Signal Processing Conference*, Sept 2013, pp. 1–5.

[22] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[23] T. Hastie and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. Springer, 2009.

[24] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[25] Y. Alwan, Z. Cvetković, and M. Curtis, "Classification of human ventricular arrhythmia in high dimensional representation spaces," *arXiv preprint arXiv:1312.5354*, 2013.

[26] M. Arafat, A. Chowdhury, and M. Hasan, "A simple time domain algorithm for the detection of ventricular fibrillation in electrocardiogram," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 1–10, 2011.

[27] S. Barro, R. Ruiz, D. Cabello, and J. Mira, "Algorithmic sequential decision-making in the frequency domain for life threatening ventricular arrhythmias and imitative artefacts: a diagnostic system," *Journal of Biomedical Engineering*, vol. 11, no. 4, pp. 320–328, 1989.

[28] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.

[29] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[30] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. B. Moody, C. Zeelenberg, and C. Marchesi, "The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography," *European Heart Journal*, vol. 13, no. 9, pp. 1164–1172, 1992.

[31] F. Nolle, F. Badura, J. Catlett, R. Bowser, and M. Sketch, "Creigard, a new concept in computerized arrhythmia monitoring systems," *Computers in Cardiology*, vol. 13, pp. 515–518, 1986.

[32] G. Moody and R. Mark, "The impact of the mit-bih arrhythmia database," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 20, no. 3, pp. 45–50, June 2001.

[33] S. Greenwald, "Development and analysis of a ventricular fibrillation detector," Master's thesis, MIT Dept. of Electrical Engineering and Computer Science, 1986.

[34] B. Raghavendra, D. Bera, A. Bopardikar, and R. Narayanan, "Cardiac arrhythmia detection using dynamic time warping of ecg beats in e-healthcare systems," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, June 2011, pp. 1–6.

[35] A. Cotter, N. Srebro, and J. Keshet, "A gpu-tailored approach for training kernelized svms," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 805–813.